

UNIVERSITY OF TECHNOLOGY SYDNEY,
AUSTRALIA.

DOCTORAL THESIS

Data Analytics and the Novice Programmer

Author:
Alireza AHADI

Supervisor:
Associate Professor
Raymond LISTER
(principal) and Dr Julia
PRIOR (co-supervisor)

*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy
in the*

Human Centred Technology Design
School of Software

January 22, 2018

Declaration of Authorship

I, Alireza AHADI, declare that this thesis titled, “Data Analytics and the Novice Programmer” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

“When I was in college, my graduation thesis was called ‘Female Directors.’ I interviewed all of the important female directors from Mexico. There were four. That was it. ”

Patricia Riggen

UNIVERSITY OF TECHNOLOGY SYDNEY, AUSTRALIA.

Abstract

Faculty of Engineering and Information Technology
School of Software

Doctor of Philosophy

Data Analytics and the Novice Programmer

by Alireza AHADI

The aptitude of students for learning how to program (henceforth *Programming learn-ability*) has always been of interest to the computer science education researcher. This issue of aptitude has been attacked by many researchers and as a result, different algorithms have been developed to quantify aptitude using different methods. Advances in online MOOC systems, automated grading systems, and programming environments with the capability of capturing data about how the novice programmer's behavior has resulted in a new stream of studying novice programmer, with a focus on data at large scale. This dissertation applies contemporary machine learning based analysis methods on such "big" data to investigate novice programmers, with a focus on novices at the early stages of their first semester. Throughout the thesis, I will demonstrate how machine learning techniques can be used to detect novices in need of assistance in the early stages of the semester. Based on the results presented in this dissertation, a new algorithm to profile novices coding aptitude is proposed and its' performance is investigated. My dissertation expands the range of exploration by considering the element of context. I argue that the differential patterns recognized among different population of novices is very sensitive to variations in data, context and language; hence validating the necessity of context-independent methods of analyzing the data.

CONFIRMATION OF ETHICS CLEARANCE

Human Negligible Low Risk Ethical clearance was granted for this PhD project by the University of Technology Sydney Human Research Ethics Committee under approval numbers ETH16-0340 (see Appendix D).

Acknowledgements

I would like to express my special appreciation and thanks to my adviser associate professor Raymond Lister, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for allowing me to grow as a research scientist. Your advice on my research have been priceless.

A special thanks to my family. Words cannot express how grateful I am to my father, mother and my beloved sister for all of the sacrifices that you've made on my behalf.

Alireza Ahadi

KEYWORDS

Novice Programmer
Human Factors
Measurement
Computer Science Education
Data Mining
Online assessment
MOOC
Databases
SQL Queries
Learning Edge Momentum
Bimodal Grade Distribution
Machine Learning
Programming Source Code Snapshot
Pattern Recognition
Classification
Supervised Machine Learning

Contents

Declaration of Authorship	iii
Abstract	vii
Acknowledgements	xi
1 Introduction	1
1.1 Motivation	1
1.2 Motivation	2
1.3 Research Questions	2
1.4 Research Design	2
1.5 Thesis Structure	3
Chapter 2: Background	3
Chapter 3: Results Overview	3
Chapter 4: Method	3
Chapter 5 to Chapter 12	3
Chapter 13: Discussion and Conclusion	3
1.6 Significance and Contribution	3
1.7 Conclusion	4
2 Background	5
2.1 Introduction	5
2.2 Who is a novice programmer?	5
2.3 Data analysis	7
2.3.1 Static success factors	8
2.3.2 Success factors and contradictory reports	10
2.3.3 Dynamically accumulating data	11
2.4 Dynamic Programming Source Code Snapshot Data Collec- tion Tools	14
2.4.1 Blackbox	14
2.4.2 CloudCoder	14
2.4.3 CodeWorkout	15
2.4.4 JS-Parsons	15
2.4.5 PCRS [118]	16
2.4.6 Problets	16
2.4.7 TestMyCode	16

2.4.8	URI Online Judge	17
2.4.9	UUhistle	17
2.4.10	Web-CAT	17
2.5	Publicly Available Programming Source Code Snapshot Datasets	18
2.5.1	Blackbox	18
2.5.2	Code Hunt	18
2.5.3	Code.Org	19
2.6	Approaches for Data Analysis	19
	Educational data mining and learning analytics . . .	19
	Machine learning and data mining approaches	20
2.7	Machine Learning in CSed	21
2.8	Assessment	23
2.9	Studies of the Novice Programming Process	24
2.10	The analysis of the novice programmer errors	24
3	Results Overview	29
3.1	Results Overview	29
3.2	The Thesis About The Thesis	29
3.2.1	Research Questions 1	30
	Paper 1. Chapter 5: Geek genes, prior knowledge, stumbling points and learning edge momen- tum: parts of the one elephant?	30
	Paper 2. Chapter 6: Exploring machine learning meth- ods to automatically identify students in need of assistance	30
	Paper 3. Chapter 7: A Quantitative Study of the Rela- tive Difficulty for Novices of Writing Seven Different Types of SQL Queries	30
	Paper 4. Chapter 8: Students' Semantic Mistakes in Writing Seven Different Types of SQL Queries	31
	Paper 5. Chapter 9: Students' Syntactic Mistakes in Writing Seven Different Types of SQL Queries and its Application to Predicting Students' Success	31
3.2.2	Research Question 2	31
	Paper 6. Chapter 10: Performance and Consistency in Learning to Program	32
	Paper 7. Chapter 11: On the Number of Attempts Stu- dents Made on Some Online Programming Exercises During Semester and their Subse- quent Performance on Final Exam Questions	32

Paper 8. Chapter 12: A Contingency Table Derived Methodology for Analyzing Course Data . . .	32
4 Method	33
4.1 Introduction	33
4.2 Background	33
4.2.1 Programming source code snapshots, collected at the University of Helsinki.	36
4.2.2 Database source code snapshots, collected at the Uni- versity of Technology Sydney.	37
5 Geek genes, prior knowledge, stumbling points and learning edge momentum: parts of the one elephant?	43
5.1 Introduction	43
5.1.1 Statement of Contribution of Co-Authors	43
5.2 PDF of the Published Paper	46
5.3 Discussion	53
6 Exploring machine learning methods to automatically identify stu- dents in need of assistance	55
6.1 Introduction	55
6.1.1 Statement of Contribution of Co-Authors	55
6.2 PDF of the Published Paper	58
6.3 Discussion	69
7 A Quantitative Study of the Relative Difficulty for Novices of Writ- ing Seven Different Types of SQL Queries	71
7.1 Introduction	71
7.1.1 Statement of Contribution of Co-Authors	71
7.2 PDF of the Published Paper	74
7.3 Discussion	81
8 Students' Semantic Mistakes in Writing Seven Different Types of SQL Queries	83
8.1 Introduction	83
8.1.1 Statement of Contribution of Co-Authors	83
8.2 PDF of the Published Paper	86
8.3 Discussion	93
9 Students' Syntactic Mistakes in Writing Seven Different Types of SQL Queries and its Application to Predicting Students' Success	95
9.1 Introduction	95
9.1.1 Statement of Contribution of Co-Authors	95
9.2 PDF of the Published Paper	98

9.3 Discussion	105
10 Performance and Consistency in Learning to Program	107
10.1 Introduction	107
10.1.1 Statement of Contribution of Co-Authors	107
10.2 PDF of the Published Paper	110
10.3 Discussion	117
11 On the Number of Attempts Students Made on Some Online Programming Exercises During Semester and their Subsequent Performance on Final Exam Questions	119
11.1 Introduction	119
11.1.1 Statement of Contribution of Co-Authors	119
11.2 PDF of the Published Paper	122
11.3 Discussion	129
12 A Contingency Table Derived Method for Analyzing Course Data	131
12.1 Introduction	131
12.1.1 Statement of Contribution of Co-Authors	131
12.2 PDF of the Submitted Paper	134
13 Discussion and Conclusion	155
13.1 Introduction	155
13.2 Overview of Research	155
13.3 Research Questions	155
13.4 Research Outcome	156
13.5 Research highlights	156
13.6 Research Significance	157
13.7 Research Findings	157
13.7.1 Data	157
13.7.2 Method	158
13.7.3 Context	158
13.8 Discussion	159
13.8.1 Data	159
13.8.2 Method	159
13.8.3 Context	160
13.9 Limitations	161
13.9.1 Data	161
13.9.2 Context	161
13.10 Recommendations	161
13.10.1 Data	162
Static data	162
Dynamic data	162

13.10.2 Method	162
13.10.3 Features of a strong algorithm to capture students learning from the source code snapshot data	162
General Attributes	162
Language Independence	163
Distribution Independence	164
Cross-Context Parameter Variation in the Construction of a Metric	164
Validity of the Operationalization	164
13.10.4 Context	165
13.11 Conclusion	165
A Definition of Authorship and Contribution to Publication	167
B Specific Contributions of Co-Authors for Thesis by Published Papers	169
C Complete list of Publications by Candidate	179
D UTS Human Ethics Approval Certificate - UTS HREC - ETH16-0340	183
E Extract from UTS Subject Outline – 31271 "Database Fundamentals" Sem. 2 2016	193
F Extract from Helsinki Subject Outline – 581325 "Introduction to Programming" Sem. 2 2016	207
G Extract from UTS Database Fundamentals – 31271, Practice Questions and Answers	215
G.1 Introduction	215
G.2 Pizza Database	215
G.2.1 Questions Used in The Practice Online SQL Test . . .	219
G.2.2 Proposed Answers for The Questions Used in The Practice Online SQL Test	219
H The Final Exam Questions Used at Helsinki University	223
I Extract from Helsinki Subject Content – 581325 "Introduction to Programming" Sem. 2 2016, Week 1.	229
Bibliography	255

List of Figures

4.1	An abstract view of the logical services where data instrumentation and collection are typically performed. Note that some data collection tools encompass multiple logical services (Adapted from Ihantola et al., 2015).	34
4.2	Data can be collected at different levels of granularity, which implies different collection frequencies and associated data set sizes (Adapted from Ihantola et al., 2015).	35
4.3	Student's home screen at AsseSQL. Students may attempt questions in any order, as many times as they wish.	38
4.4	Screen for an individual question. The question and the simple output to clarify the question are presented for each question in the test.	39
4.5	Feedback screen for an individual question with the output of model answer as well as the student's answer are presented in this page.	39
4.6	Back to the student's home screen: students may return to this at anytime.	40
4.7	A sample of the ERD used in the test.	41

List of Tables

3.1	Research Questions, the thesis about the thesis, and the corresponding chapters	29
5.1	Authors' Area of Contribution for The Paper Corresponding to Chapter 5	44
6.1	Authors' Area of Contribution for The Paper Corresponding to Chapter 6	56
7.1	Authors' Area of Contribution for The Paper Corresponding to Chapter 7	72
8.1	Authors' Area of Contribution for The Paper Corresponding to Chapter 8	84
9.1	Authors' Area of Contribution for The Paper Corresponding to Chapter 9	96
10.1	Authors' Area of Contribution for The Paper Corresponding to Chapter 10	108
11.1	Authors' Area of Contribution for The Paper Corresponding to Chapter 11	120
12.1	Authors' Area of Contribution for The Paper Corresponding to Chapter 12	132
G.1	List of questions and their corresponding covered topic. . . .	220
G.2	Proposed Answers for the questions presented in G.1	221

List of Abbreviations

ACC	Accuracy
CMS	Course Management System
CTP	Computational Thinking Patterns
DM	Data Mining
DT	Decision Tree
EDM	Educational Data Mining
EQ	Error Quotient
FN	False Negative
FP	False Positive
FDR	False Discovery Ratio
FNR	False Negative Ratio
FPR	False Positive Ratio
IDE	Integrated Development Environment
JAR	Java Archive File
LMS	Learning Management System
LN	Logical Necessity
LS	Logical Sufficiency
LSI	Kolb's Learning Style Inventory
ML	Machine Learning
MSLQ	Motivated Strategies Learning Questionnaire
NN	Neural Network
NPSM	Normalized Programming State Model
PCA	Principal Component Analysis
RF	Random Forest
RED	Repeated Error Density
RSE	Rosenberg Self-Esteem
SEN	Sensitivity
SPC	Specificity
SQL	Structured Query Language
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
WATWIN	WATson & GodWIN
Web-CAT	Web-based Center for Automated Testing

List of Publications by Candidate

Below are the publications first authored by the Candidate which contribute to this thesis by publication.

1. Ahadi, A. and Lister, R., 2013, August. Geek genes, prior knowledge, stumbling points and learning edge momentum: parts of the one elephant?. In Proceedings of the ninth annual international ACM conference on International computing education research (pp. 123-128). ACM. (See Chapter 5)
2. Ahadi, A., Prior, J., Behbood, V. and Lister, R., 2015, June. A Quantitative Study of the Relative Difficulty for Novices of Writing Seven Different Types of SQL Queries. In Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education (pp. 201-206). ACM. (See Chapter 7)
3. Ahadi, A., Lister, R., Haapala, H. and Vihavainen, A., 2015, July. Exploring machine learning methods to automatically identify students in need of assistance. In Proceedings of the eleventh annual International Conference on International Computing Education Research (pp. 121-130). ACM. (See Chapter 6)
4. Ahadi, A., Behbood, V., Vihavainen, A., Prior, J. and Lister, R., 2016, February. Students' Syntactic Mistakes in Writing Seven Different Types of SQL Queries and its Application to Predicting Students' Success. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education (pp. 401-406). ACM. (See Chapter 9)
5. Ahadi, A., Prior, J., Behbood, V. and Lister, R., 2016, July. Students' Semantic Mistakes in Writing Seven Different Types of SQL Queries. In Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education (pp. 272-277). ACM. (See Chapter 8)
6. Ahadi, A., Lister, R. and Vihavainen, A., 2016, July. On the Number of Attempts Students Made on Some Online Programming Exercises During Semester and their Subsequent Performance on Final Exam Questions. In Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education (pp. 218-223). ACM. (See Chapter 11)
7. Ahadi, A., Lal, S., Leinonen, J., Lister, R. and Hellas, A. Performance and Consistency in Learning to Program. Australian Computing Education Conference, 2017 (Award winning paper for best student paper). (See Chapter 10)

8. Ahadi, A., Lister, R., Hellas, A. A Contingency Table Derived Methodology for Analyzing Course Data. 17, 3, (August 2017), 19 pages.
DOI: <https://doi.org/10.1145/3123814> (See Chapter 12)

THESIS BY PUBLICATION

This thesis is presented in the format of scholarly papers published during the period of my candidature, according to UTS regulations set out in the website <http://uts.edu.au/current-students/dab/uts-graduate-research-school>. The papers included in this thesis by publication form a research narrative which is summarized in Chapter 5. Each paper then forms a separate chapter of this thesis, inserted in its published format.

To all novice programmers...

